Contribution to

The IEEE 802.3 100G-EPON Task Force Meeting, Mar. 14-16, 2016

Bonding requirements for 100G-EPON

Frank Effenberger, Duane Remein

Fixed Access Network Research

Futurewei Technologies, Huawei R&D USA

www.huawei.com

HUAWEI

Introduction

- From early in the project, there has been a stated desire to support “100G MAC rates”
- With the assumption that the per channel rate is 25G, this naturally raises the issue of combining (bonding) the channels
- In the case of EPON, bonding at a lower layer presents many architectural issues – last we heard there was no solution!

- This presentation considers the real need for bonding, revealing some of the old assumptions may no longer be true

What's the application of 100G service?

- If one sells a 100G interface to a customer, what is he liable to use it for?
  - We certainly invite the participating operators to tell us!
- One "poster child" application is a data back-up system, that does a massive data dump every night
  - The story goes, such an application would push 100G of data across the network on a single TCP/IP port/address
- This is actually not true!
  - No computer can realistically fill a 100G interface, and in fact most can't even fill a 10G link
  - Large capacity applications are heavy users of parallel computing, where tasks are spread over multiple processors
  - So, at a hardware level, the 100G 'flow' is already being disaggregated

High bandwidth applications

- Looking deeper, modern high bandwidth applications are already solving this problem at several levels
  - It is a well known problem that the modern network is distributed and disaggregated – If you want to play at this level, you must know how to deal with this
  - The application itself is likely to do some level of parallel computing / task dispatching to use resources more efficiently
  - Common operating systems (e.g., Linux) also implement a version of link-aggregation (NIC agg), to make it easier
  - The essence of cloud computing is to separate the abstract view of the network from the actual implementation of the network
- Are we chasing a ghost?

What about link aggregation?

- The standard solution is link aggregation (IEEE 802.1AX)
  - There are several operating modes
- The most common mode is Balanced-XOR
  - Packets are routed based on a hash of header contents
  - Flows are guaranteed to be in-order, but links may have poor balance
- To address imbalance, there is Adaptive Load Balancing
  - Flows are routed to the interfaces based on their current load
  - Per flow packet ordering is still ensured, and balance is better
- If packet reordering is ok, there is round robin mode
  - Packets are distributed over the channels in a trivial RR scheduling method

Is packet reordering a sin?

- At present, it seems flow-based LAG will work well
  - There is a wide range of solutions already
  - The application and OS can get involved to make it work
- But what if we are stuck with a “difficult case”?
  - This is where the round robin mode would be used
- Round robin mode will introduce a certain amount of packet reordering, driven by variations in packet sizes and link speeds
  - Notably, in our case, we’re aggregating channels that are all the same speed (25G), and so our reordering is bounded
- The problem with packet reordering sits with TCP
  - Early implementations treated misordered packets as evidence of packet loss
  - This is no longer true, and certainly any system that is aiming to push 25+G of bandwidth will be using a more modern TCP
- Bottom line: TCP today can tolerate some reasonable level of packet reordering

Impact to 100G-EPON

- Note the previous slide stated that if the link bandwidths are close in bandwidth, then even round robin LAG can solve the worst cases
- In the PON downstream, the channels will be equal – no problem there
- In the upstream, we could have a problem
  - The OLT needs to implement a DBA that provides equal bandwidth to all the links that a particular ONU is using
  - Otherwise, the round robin distribution of traffic will produce significant misordering
- This raises an interesting point: how should a 100G ONU be given 40G of bandwidth?
  - 4x10G, 2x20G, or 25+15G
  - The first two will work well, avoiding packet misordering
  - The last one will cause serious packet misordering

Actually, normal DBA works

- Imagine we have a 100G ONU that is operating on 4 channels, and it distributes its packets using round robin
  - The upstream buffers will be more or less equal
  - The OLT will respond to that by giving more or less equal grants to all four channels
  - The bandwidth will be balanced naturally
- Suppose that for power saving reasons, the ONU is commanded to only operate on 2 channels
  - Traffic would go into only those two buffers
  - Those two channels would be given twice as much as before, but still more or less equal
- Good news: nothing is broken

Thank you

www.huawei.com